Exascale in 2020 - Which components might be relevant for the CERN experiments and the LHC Computing Grid?

Sverre Jarp
CERN openlab CTO
Intel Exascale Meeting, Nice,
23rd October 2013

Accelerating Science and Innovation

Acknowledgement

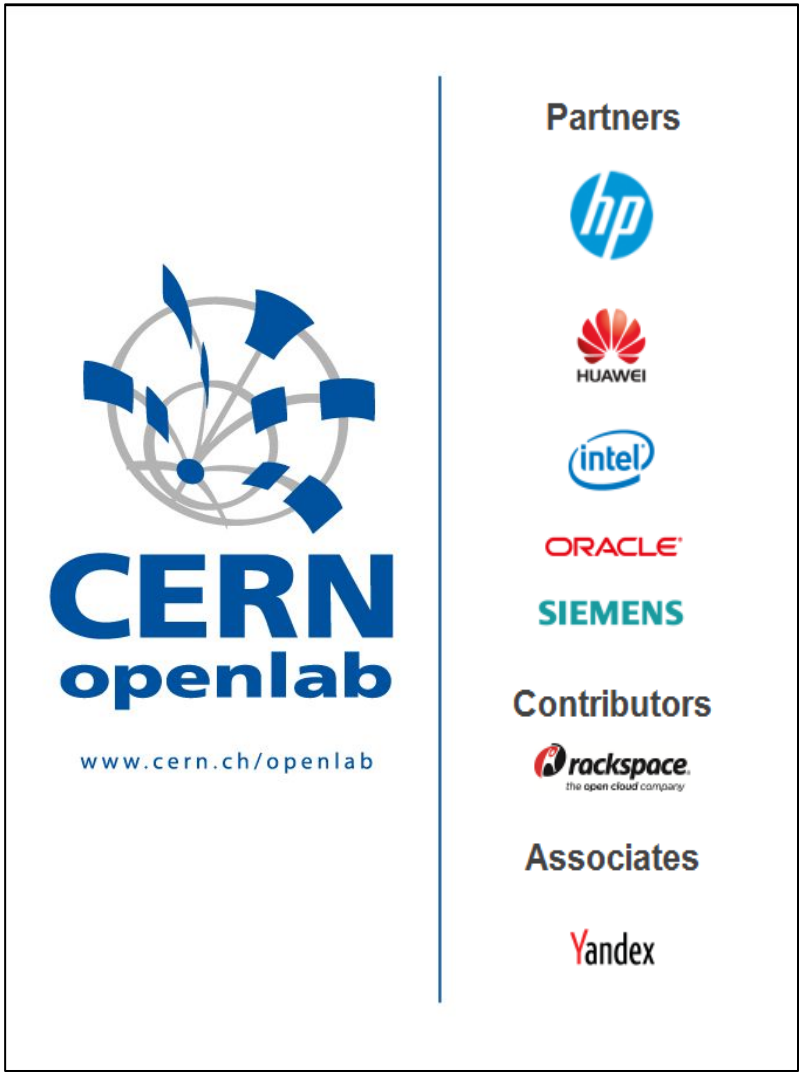
- I am indebted to many colleagues who helped produce these slides:
- Alberto Pace (IT/DSS)
- Ian Bird (WLCG Project leader)
- Alberto di Meglio and Andrzej Nowak (CERN openlab CTO office)
- Niko Neufeld (LHCb)
- Philippe Canal (CHEP 2013)
- Torre Wenaus (CHEP 2013)
- … and many additional colleagues in HEP (High Energy Physics) Computing

What is CERN ?

- The European Particle Physics Laboratory based in Geneva, Switzerland
  - Current accelerator: The Large Hadron Collider (LHC)
- Founded in 1954 by 12 countries for fundamental physics research in a post-war Europe
- Today, it is a global effort of 20 member countries and scientists from 110 nationalities, working on the world's most ambitious physics experiments
- ~2'300 personnel, > 10'000 users
- ~900 million € yearly budget

CERN openlab

- **A unique research partnership between CERN and the industry**
- **Objective: The advancement of cutting-edge computing solutions to be used by the World-wide LHC Computing Grid**
- **First phase started in 2003**
- **Discussions are starting for Phase V (2015 – 2017)**

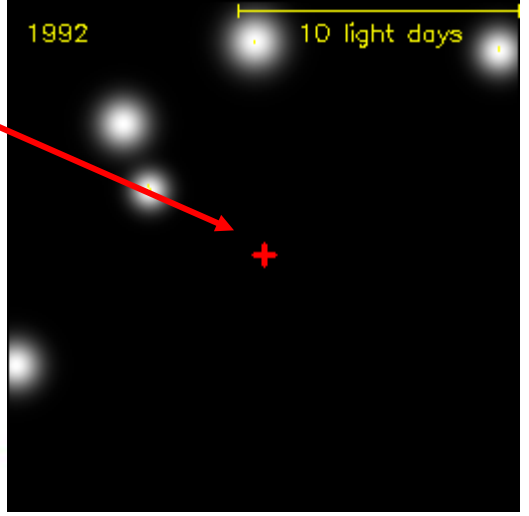


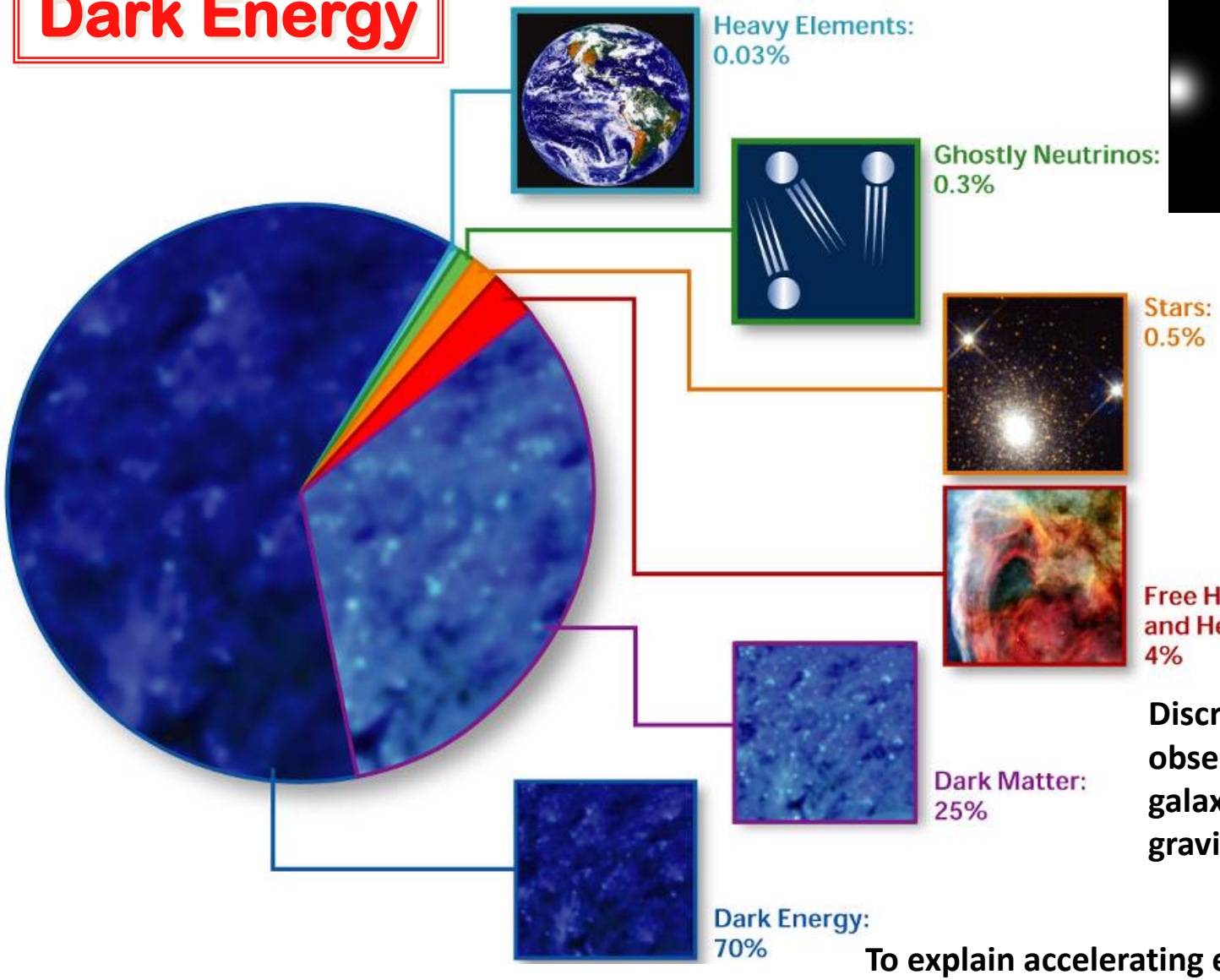
Why do we need a “CERN”?

Especially after the Higgs boson was recently added as the last piece to the Standard Model

The focus is now on Dark Matter and Dark Energy

Black hole





> 95% UNKNOWN STUFF IN THE UNIVERSE

Discrepancy between observed mass of stars & galaxies and their gravitational motion

To explain accelerating expansion of the universe

So, how do you get from this

to this

Higgs boson-like particle discovery claimed at LHC

COMMENTS (1665)

By Paul Rincon

Science editor, BBC News website, Geneva

The moment when Cern director Rolf Heuer confirmed the Higgs results

Cern scientists reporting from the Large Hadron Collider (LHC) have claimed the discovery of a new particle consistent with the Higgs boson.

Relat

and this

Physics 2013

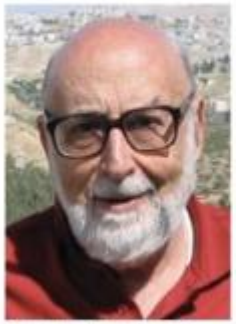


Photo: Pnicolet via Wikimedia Commons

François Englert

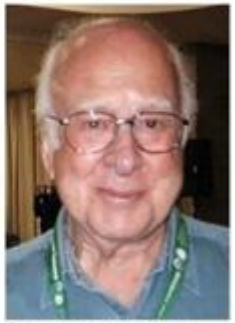
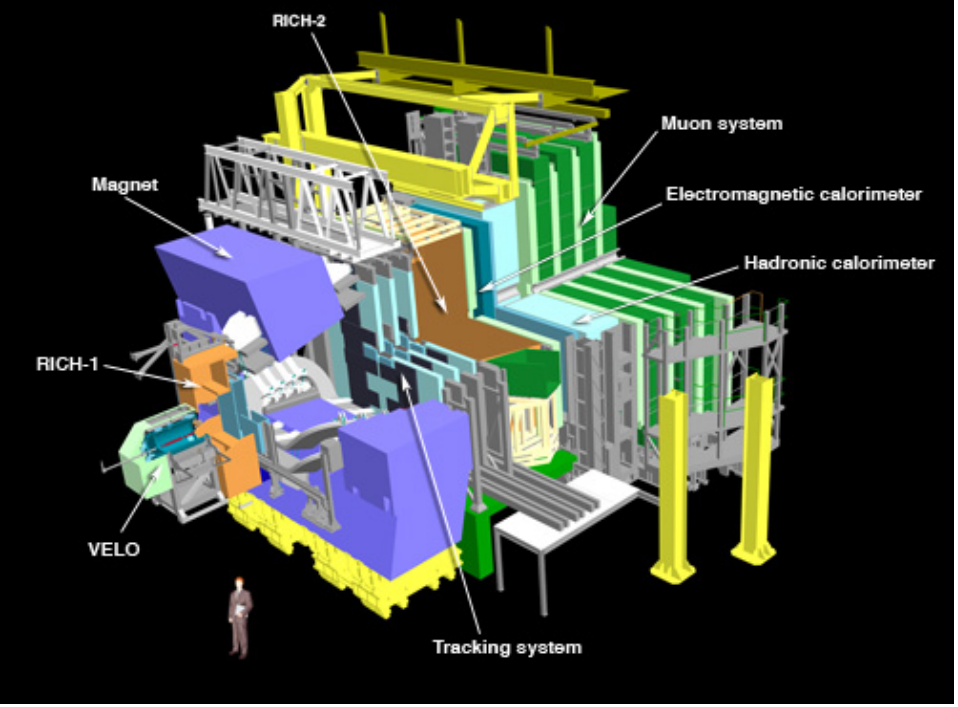
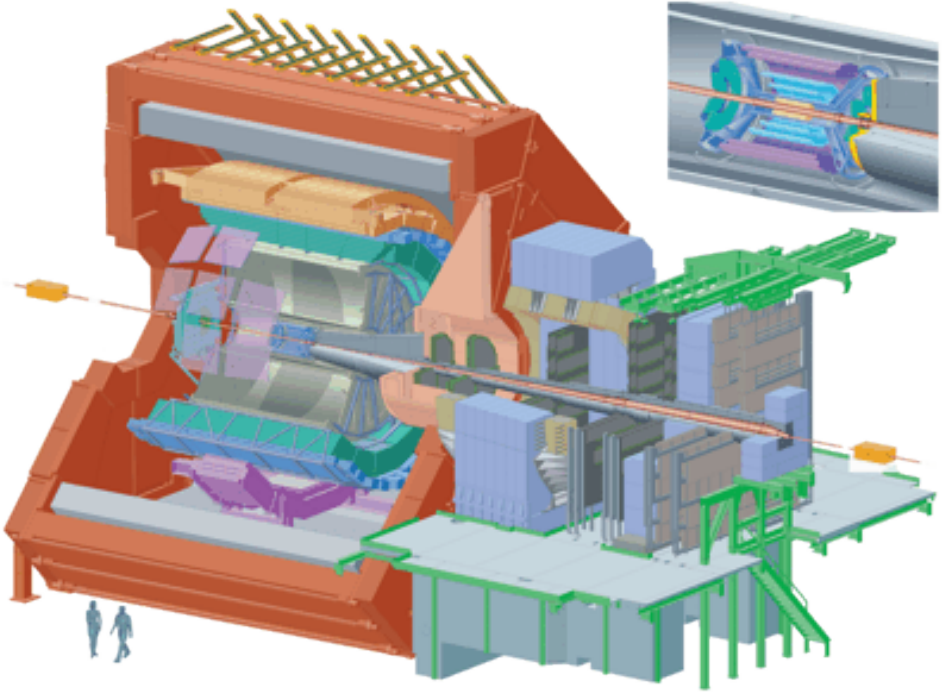


Photo: G-M Greuel via Wikimedia Commons

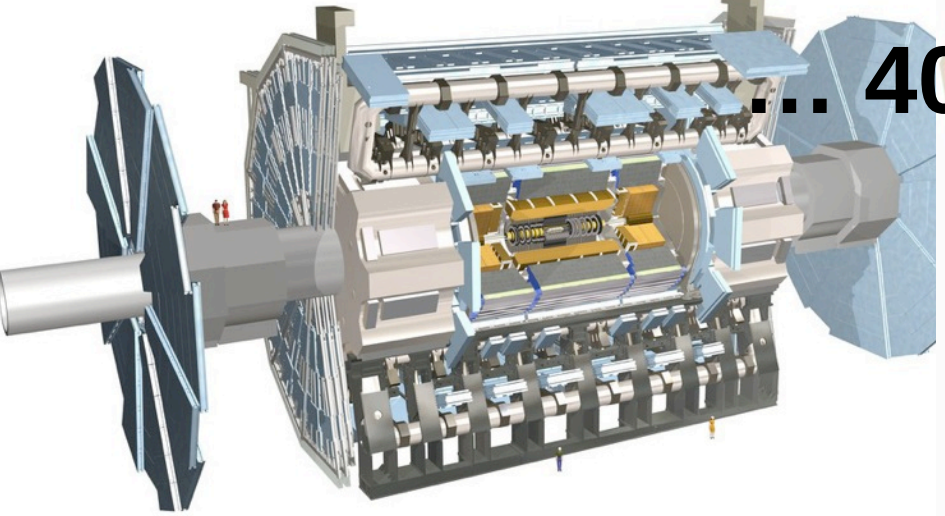
Peter W. Higgs

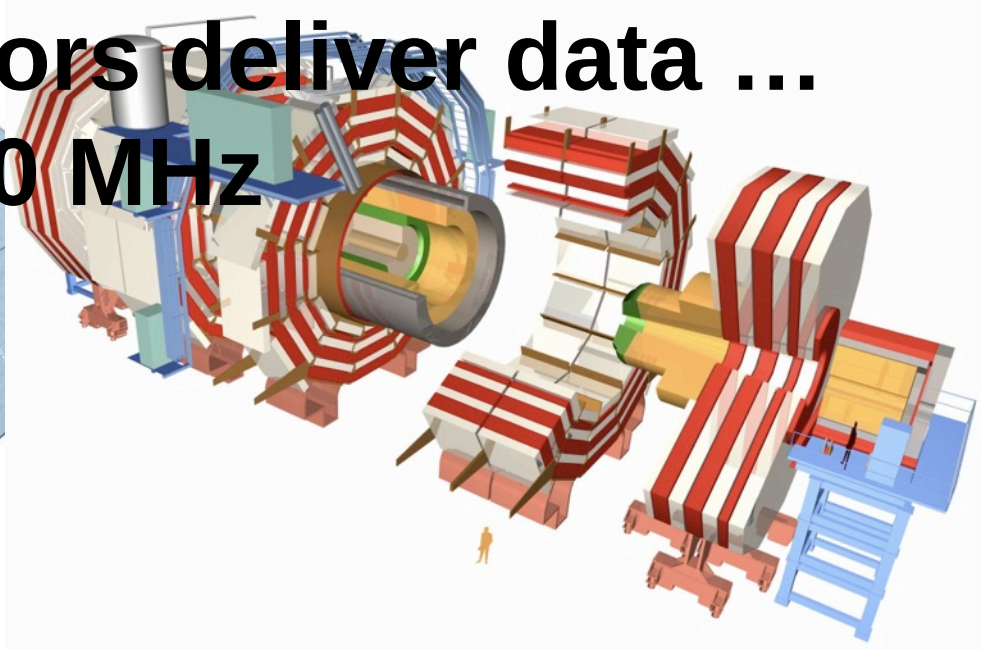




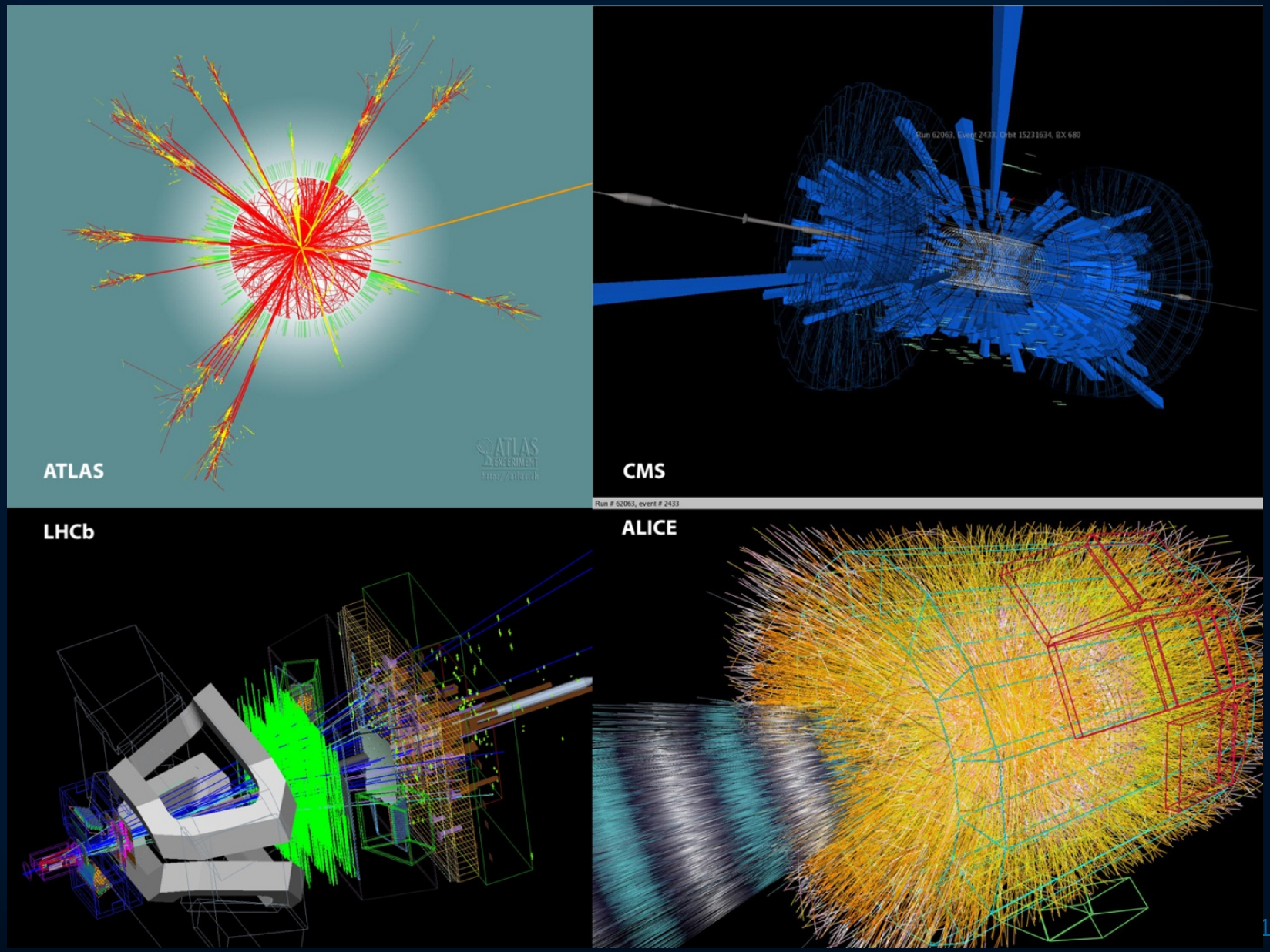
150 million sensors deliver data …

… 40 MHz





What do events look like?



Some history of scale...

Date	Collaboration sizes	Data volume, archive technology
Late 1950's	2-3	Kilobits, paper notebooks
1960's	10-15	KB, punchcards
	~35	MB, tape
	~100	GB, tape, disk
	700-800	TB, tape, disk
2010's	~3000	PB → EB, tape, disk

Rubbia's Nobel Prize (1984)

For comparison:
1990's: Total LEP data set ~few TB
Would fit on 1 tape today

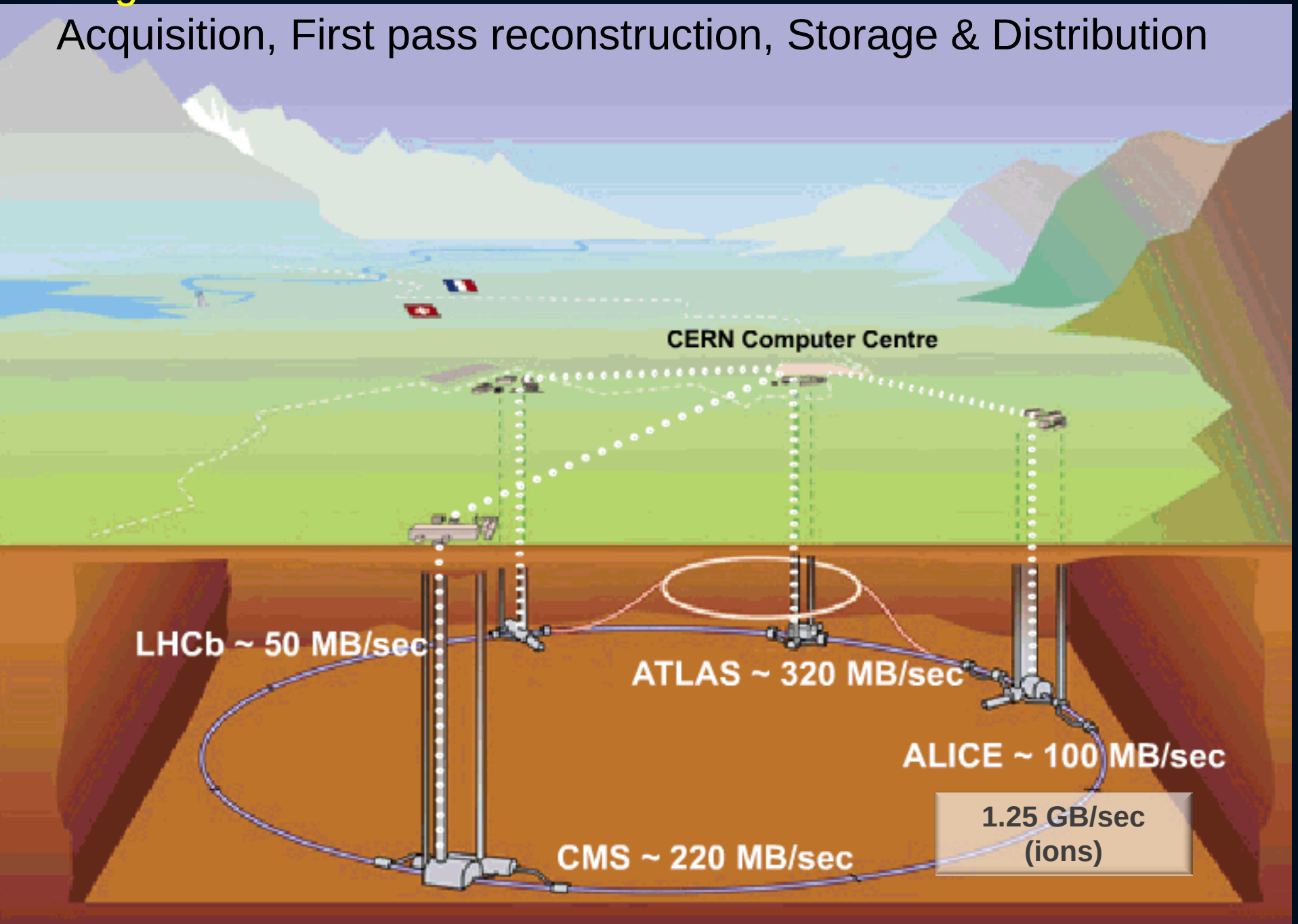
Today: 1 year of LHC data ~15 PB

CERN has about 80,000 physical disks to provide in the order of 100 PB of reliable storage (30 PB in Tier0)

Run 1 (2009 – 2012)

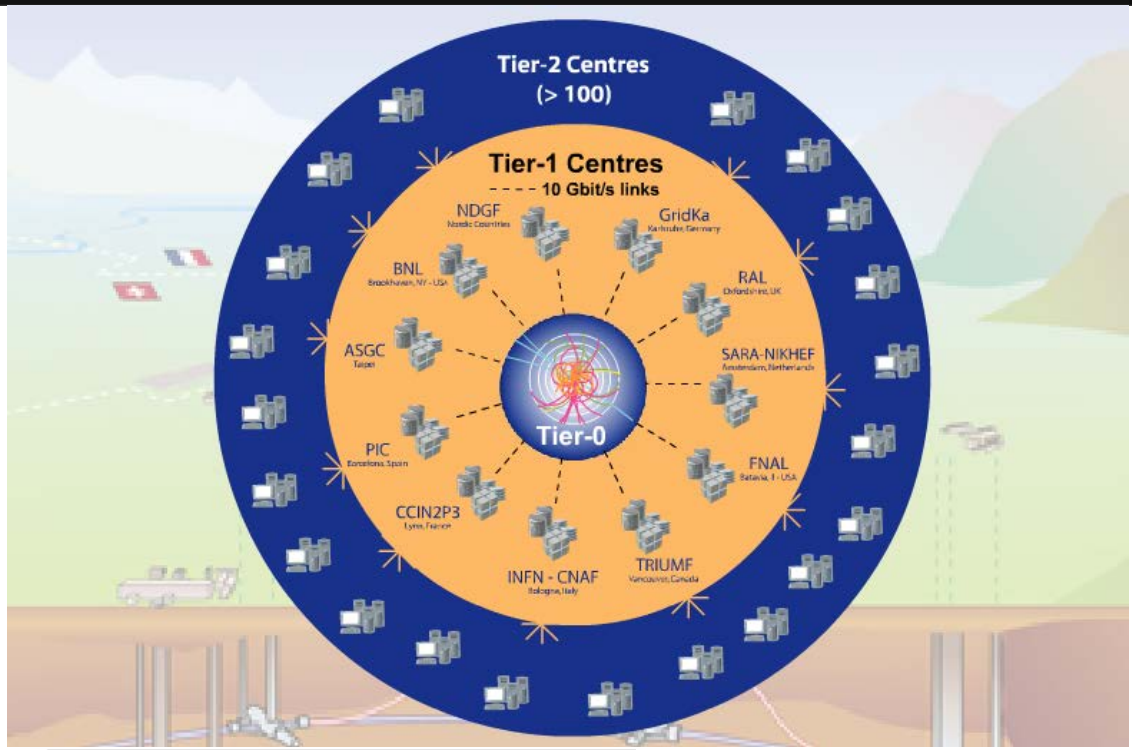
Original rates for the Tier 0 at CERN:

Acquisition, First pass reconstruction, Storage & Distribution



World-wide LHC Computing Grid

- A distributed computing infrastructure to provide the production and analysis environments for the LHC experiments
- Managed and operated by a worldwide collaboration between the experiments and the participating computer centres
- The resources are distributed – for funding and sociological reasons
- Our task was to make use of the resources available to us – no matter where they are located

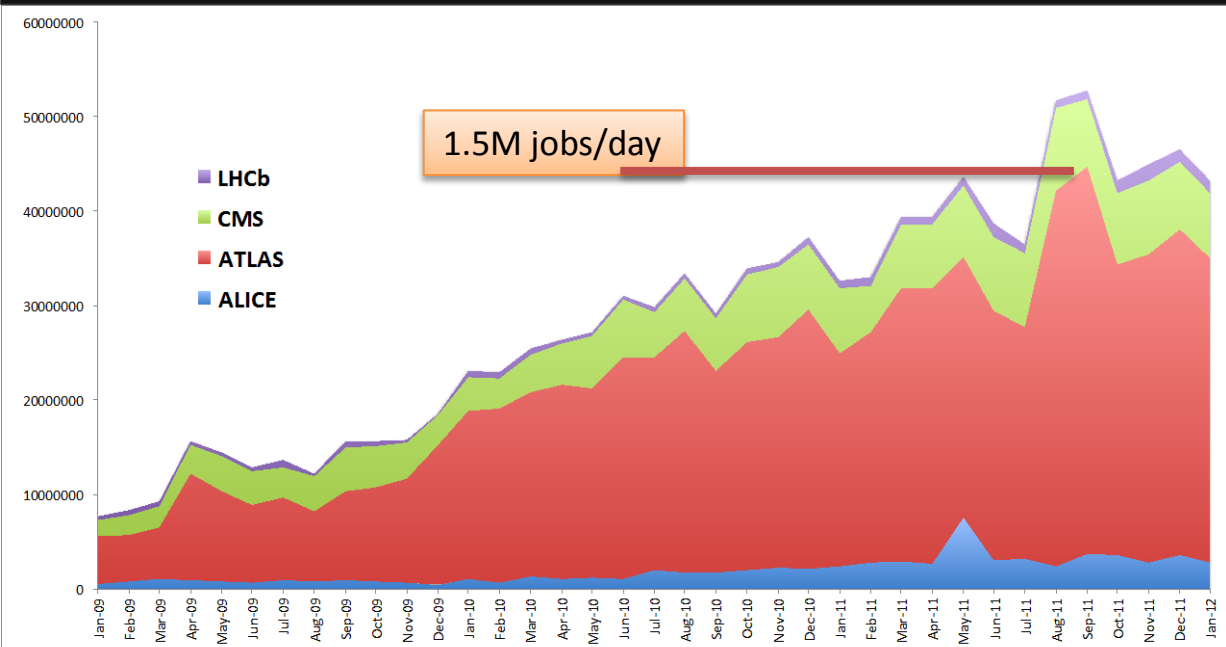


Tier-0 (CERN):
- Data recording
- Permanent storage
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (~130 centres):
- Simulation
- End-user analysis

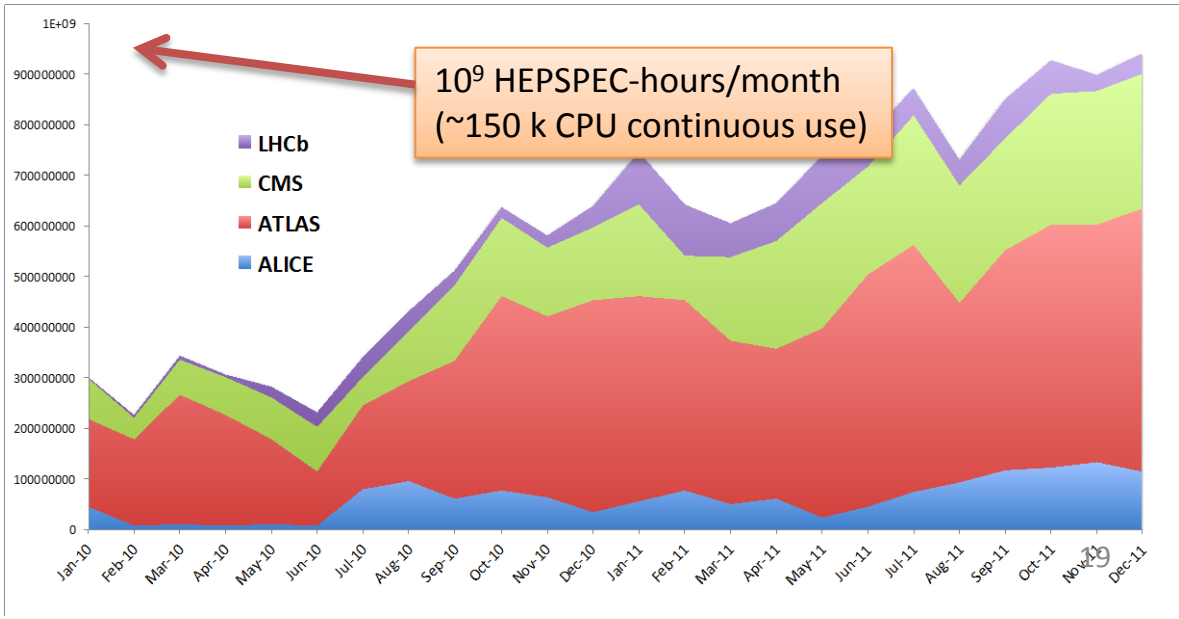
Processing on the grid



Usage continues to grow…

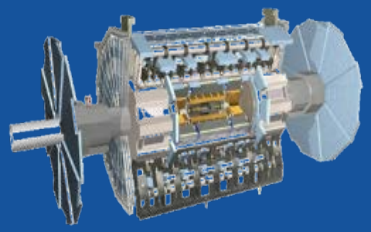
- # jobs/day
- CPU usage

~ 150,000 years of CPU delivered each year



This is close to full capacity
We always need more!

Data Handling and Computation



Online Triggers and Filters

Offline Reconstruction

Selection & reconstruction

Event reprocessing

100% (raw data, 6 GB/s)

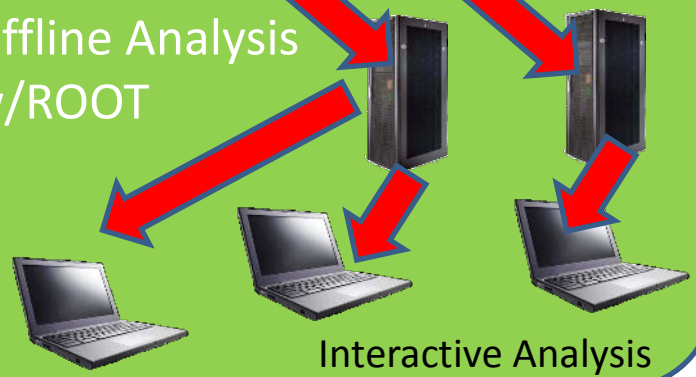
10% (event summary)

Processed data (active tapes)

Batch Physics Analysis

1%

Offline Analysis w/ROOT



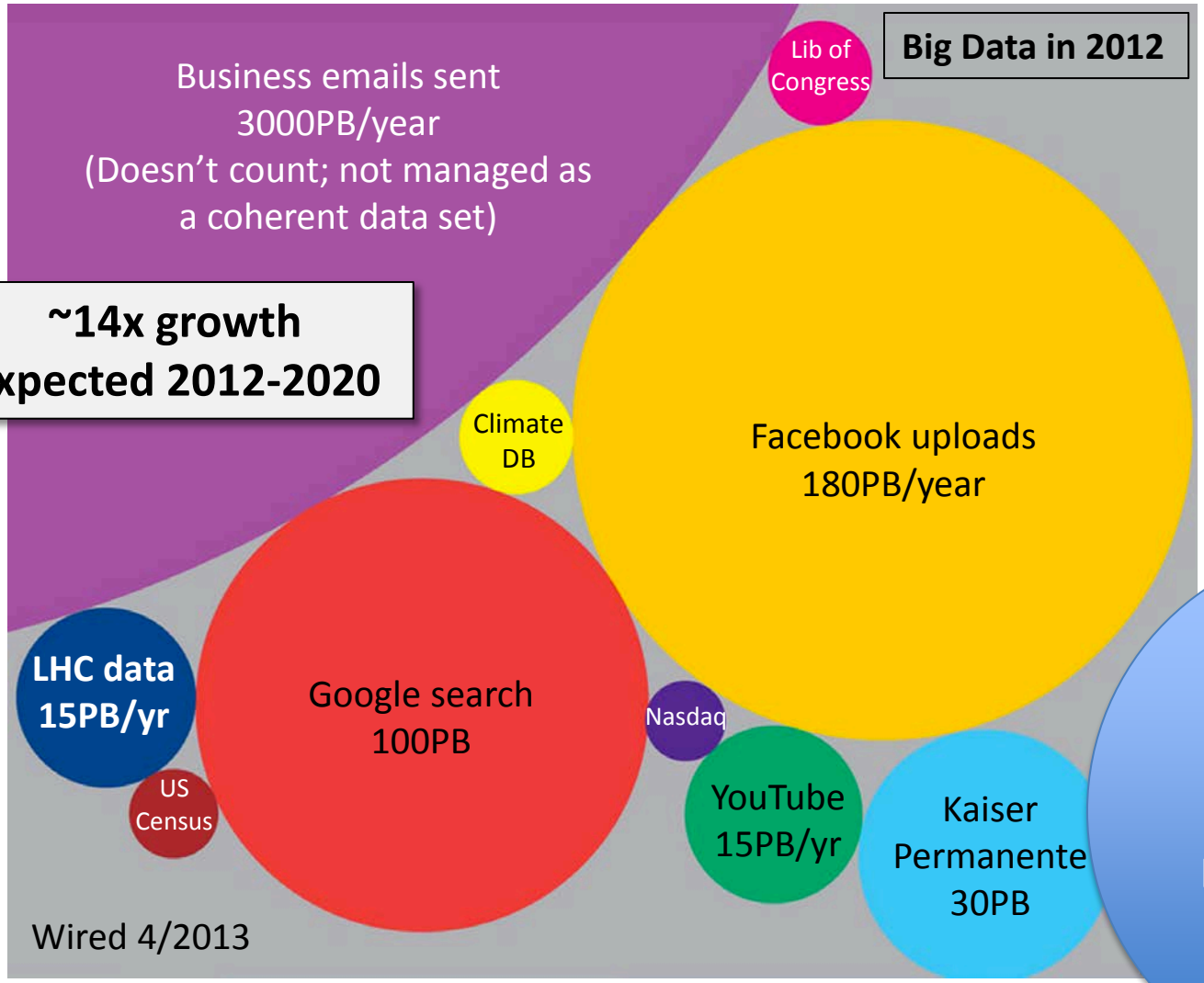
Interactive Analysis

Event simulation

Geant 4

Offline Simulation

Data Management
Where is LHC in Big Data Terms?



In 2012: 2800 exabytes created or replicated
1 Exabyte = 1000 PB

Reputed capacity of NSA's new Utah center: 5000 ExaB (50-100 MW)

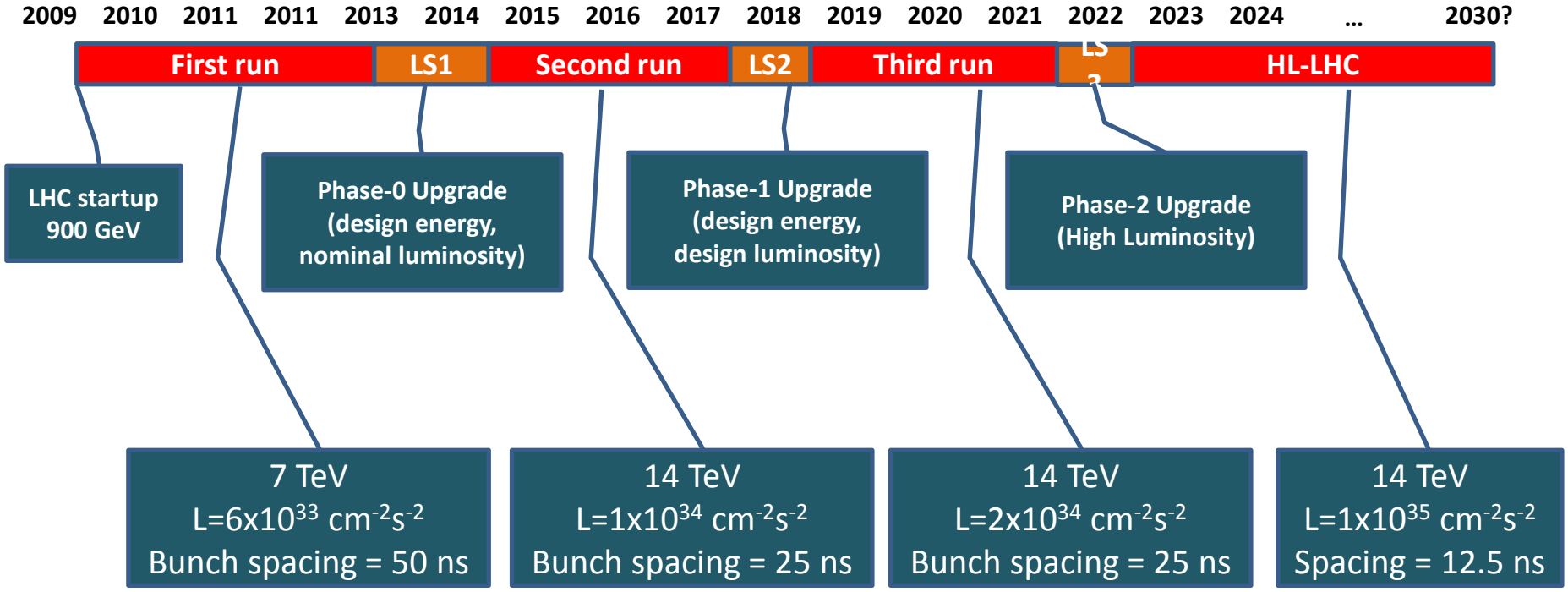
Current ATLAS data set, all data products: 140 PB

http://www.wired.com/magazine/2013/04/bigdata/

The LHC Road Map

LHC Beyond 2015

to the High Luminosity LHC (HL-LHC)



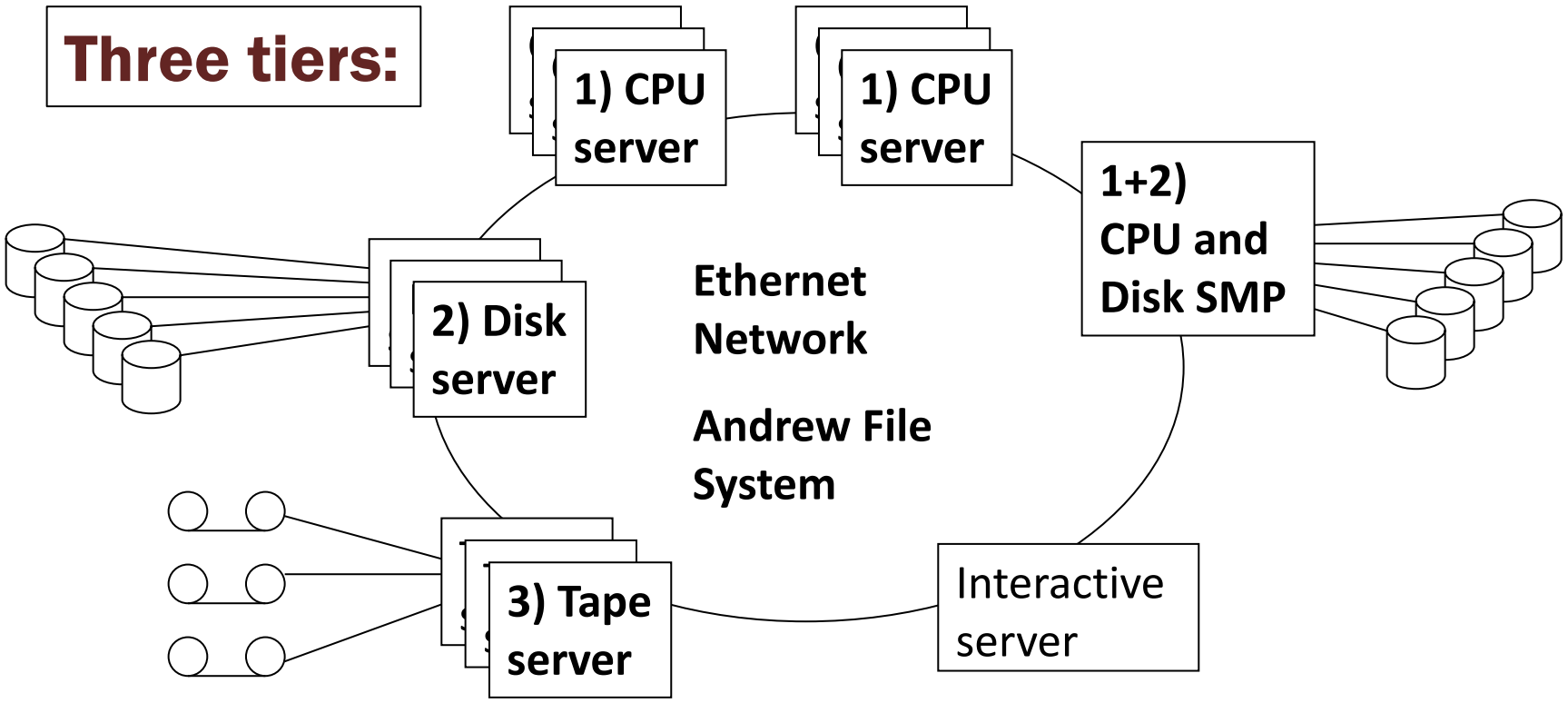
- **Trigger rates, event complexity increase steadily through machine and detector upgrades**
  - **~15 PB/year LHC raw data now; ~130 PB/year in 2021 (end of Run 3)**
  - **Very rough estimate for new raw data per year in Run 4: 400 PB**

Computer Centre (Off-line Computing)

SHIFT architecture

(Scalable Heterogeneous Integrated Facility)

- **A versatile scale-out architecture**
- **In 2001 it won the 21st Century Achievement Award issued by Computerworld**



Active tapes

- **Inside a huge storage hierarchy tapes may be advantageous!**

We use tape storage products from multiple vendors

Exascale in 2020

- **Exascale Targets compared to DOE's current HPC systems (HPCwire, 16 Sept. 2013)**

	Mid-2013	2020
System peak (Linpack)	~25 PF	1000 PF
System power (MW)	8 - 9	~20
System memory (PB)	~1	~64
Storage (PB)	10 - 15	500 - 1000
MTTI (days)	~7	~1

2013 comparison

- **Translated into Sandy Bridge cores**

	DOE	WLCG	Comparison
System peak (Linpack)	~25 PF	N.A.	
Achievable GF/core (SNB)	20		
Cores	$1.25 * 10^6$	$0.35 * 10^6$	Smaller
System memory (PB)	~1	~1	Similar
Storage (PB)	10 – 15	150	Bigger
MTTI (days)	~7	N.A.	Much better

2020 comparison

- **Needs to be taken with a pinch of salt**

	DOE	WLCG	Comparison
System peak (Linpack)	1000 PF	N.A.	
Required CPU growth	ML^2	ML	Much smaller
System power (MW)	20 (?)	25 – 30	Similar
System memory (PB)	~64	~64	Similar
Storage (PB)	500 – 1000	2000 - 3000	Bigger
MTTI (days)	~1	N.A.	Much better

ML (Moore's Law)

CERN (Combined Meyrin/Geneva and Wigner/Budapest)

- Available power: 3.5 MW + 2.0 MW
- Interconnect: Two links at 100 Gb/s
- Wigner Centre in full operations next year
- Currently in Meyrin:
  - 10’000 EP servers (90’000 cores)
  - LAN interconnect: 1 and 10 Gb/s
  - ~100 PB disk space
  - >100 PB tape storage
- Evolution:
  - CPU capacity: Somewhat higher than ML
    - But, how much more?
  - CPU cores: Whichever exhibit the best Perf/€/W
  - System memories: Remain opportunistic
  - LAN interconnect: 10 and 100 Gb/s
  - Storage: Multiple exabytes (both disks and tapes)
    - Need to cater for explosion in data from experiments

Experiments
(On-line Computing)

Online Trigger Farms in Run 1

	ALICE	ATLAS	CMS	LHCb
# cores (+ hyperthreading)	2700	17'000	13'200	15'500
# servers (mainboards)	~ 500	~ 2000	~ 1300	1574
total available cooling power [kW]	~ 500	~ 820	800	525
total available rack-space (Us)	~ 2000	2400	~ 3600	2200

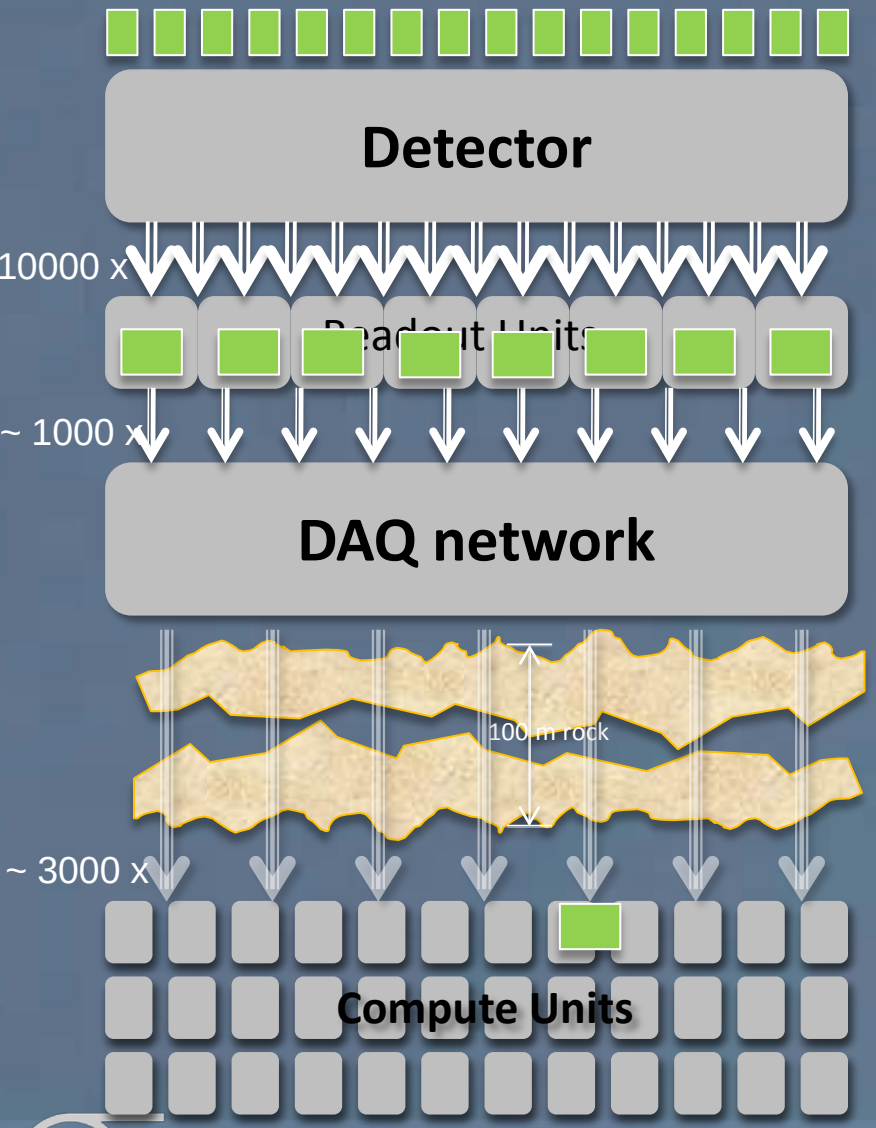
Massive upgrades foreseen for Run 2

Adding more intelligence into the Data Acquisition Systems in the future

	Event-size [kB]	Rate of events into HLT [kHz]	HLT (High Level Trigger) bandwidth [Gb/s]	Year
ALICE	20’000	50	8000	2019
ATLAS	4000	200	6400	2022
CMS	4000	1000	32’000	2022
LHCb	100	40’000	32’000	2019

The experiments will move into the Terabit range. Two of them (ALICE and LHCb) already in 2019

Data Acquisition (generic example)



Every Readout Unit has a piece of the collision data
All pieces must be brought together into a single compute unit
The Compute Unit runs the software filtering (High Level Trigger – HLT)

GBT (GigaBit Transceiver): custom radiation- hard link from the detector 3.2 Gbit/s

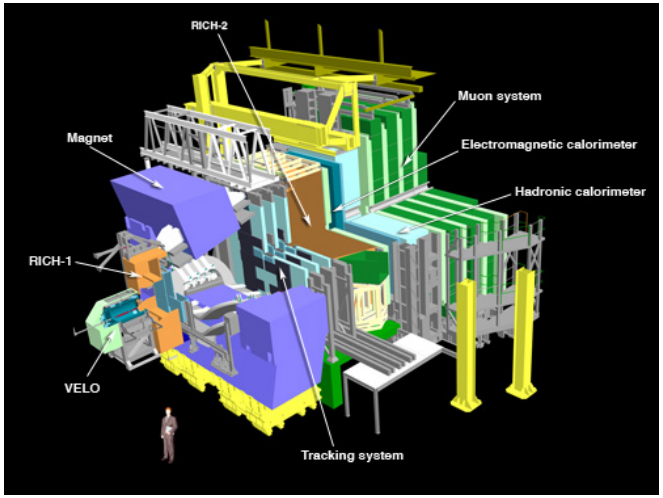
DAQ (“event-building”) links – some LAN (10/40/100 GbE / InfiniBand)

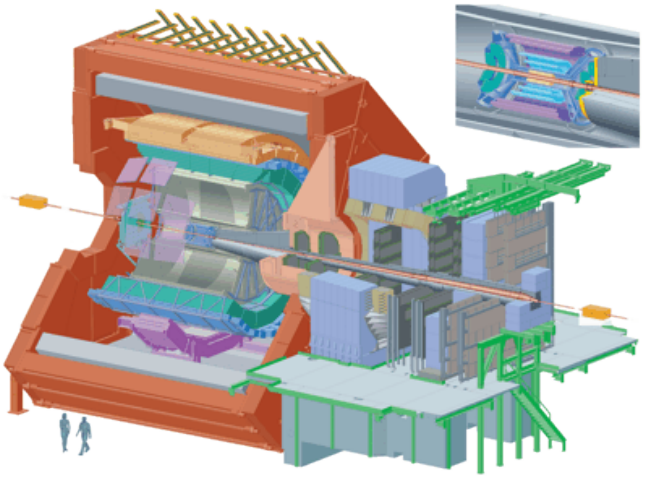
Links into compute-units: typically 10 Gbit/s (because filtering is currently compute-limited)

A closer look at ALICE and LHCb

- **Planned capacities for 2019:**

	ALICE	LHCb
Data Rate (HLT in Gb/s)	8’000	32’000
Network speed (Gb/s)	10 - 100	10 - 100
Processing power increase	100x	40x

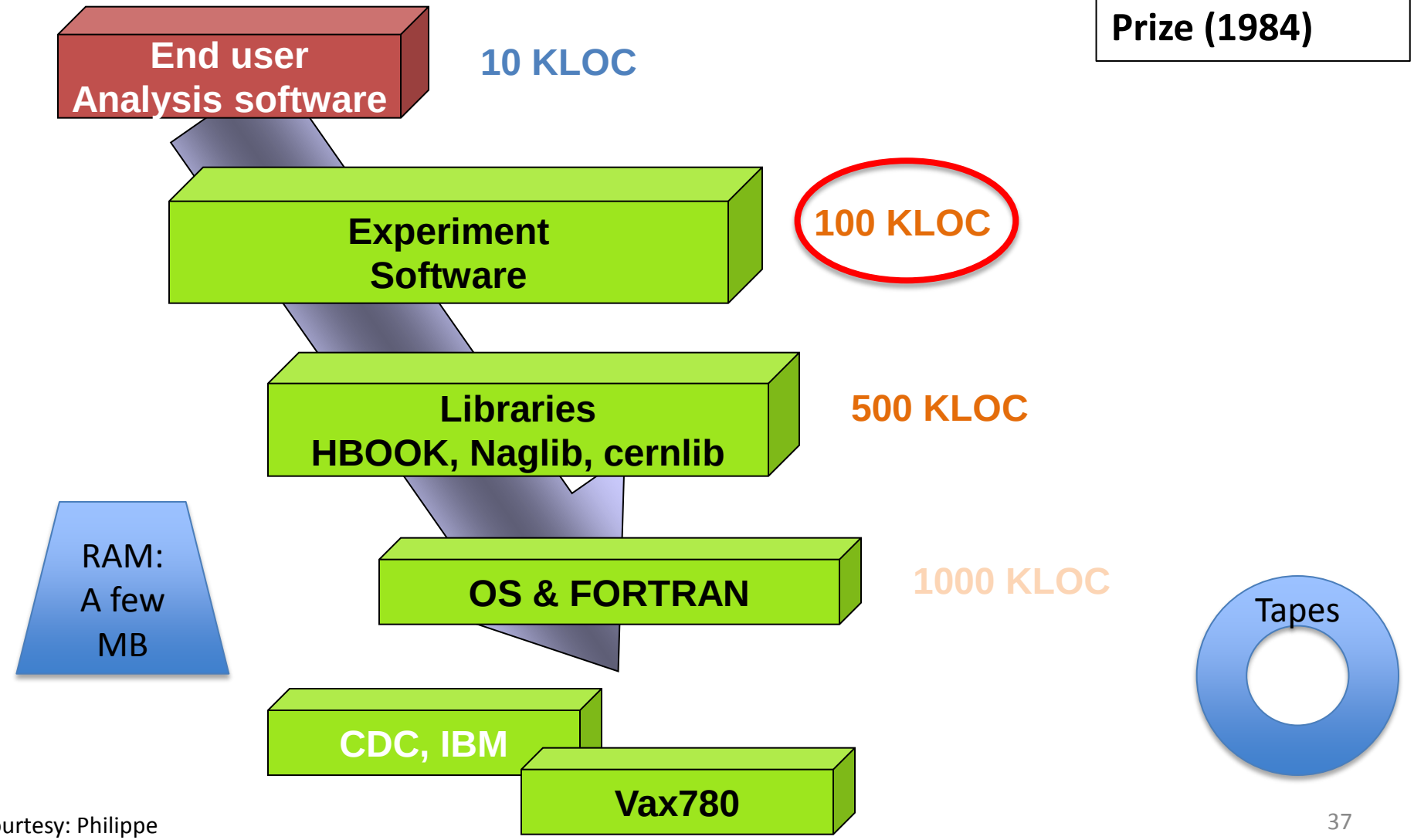




Future software.
Is this the real problem?

Systems in 1980

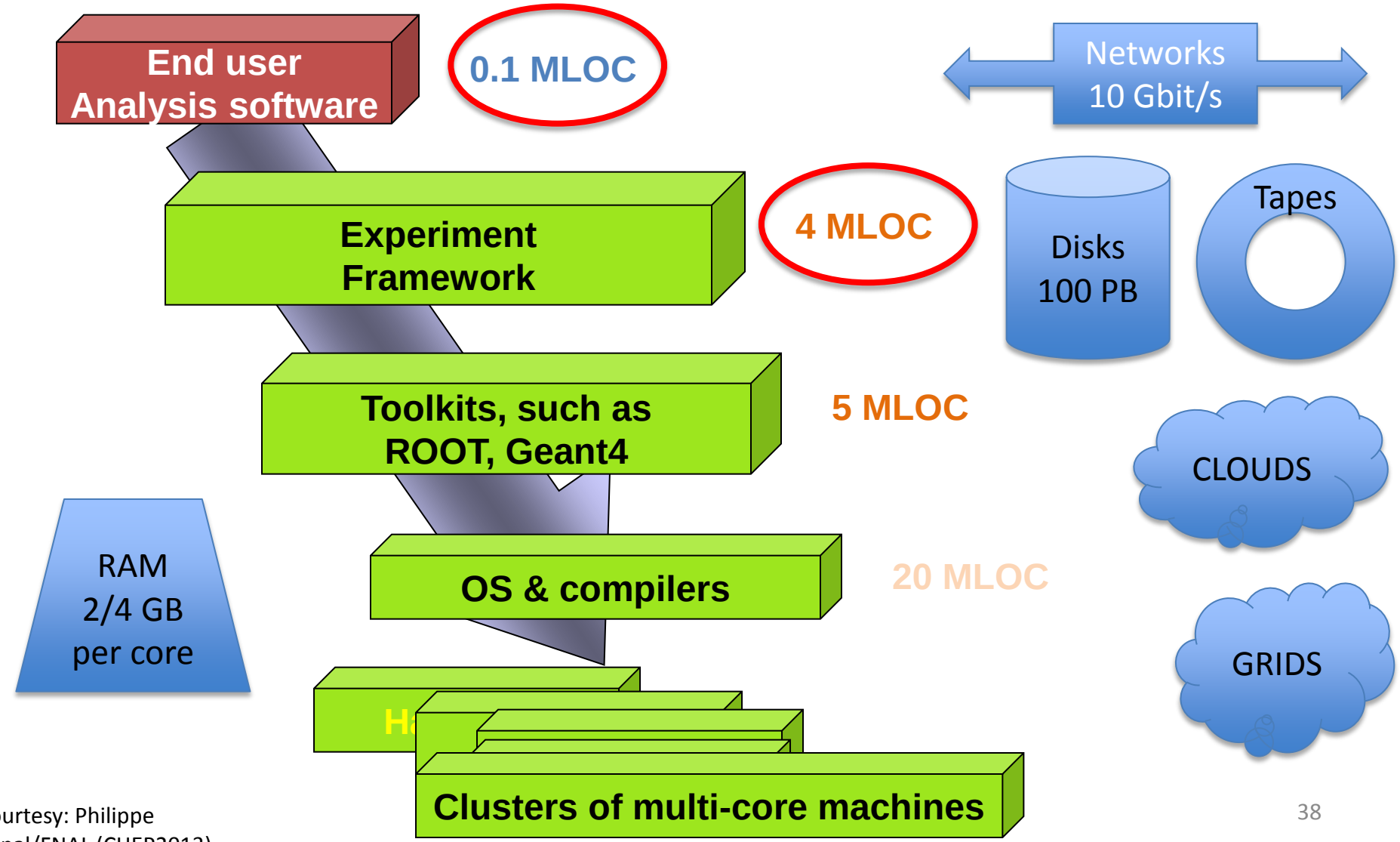
The computing environment at the time of Rubbia's Nobel Prize (1984)



Courtesy: Philippe Canal/FNAL (CHEP2013)

Systems today

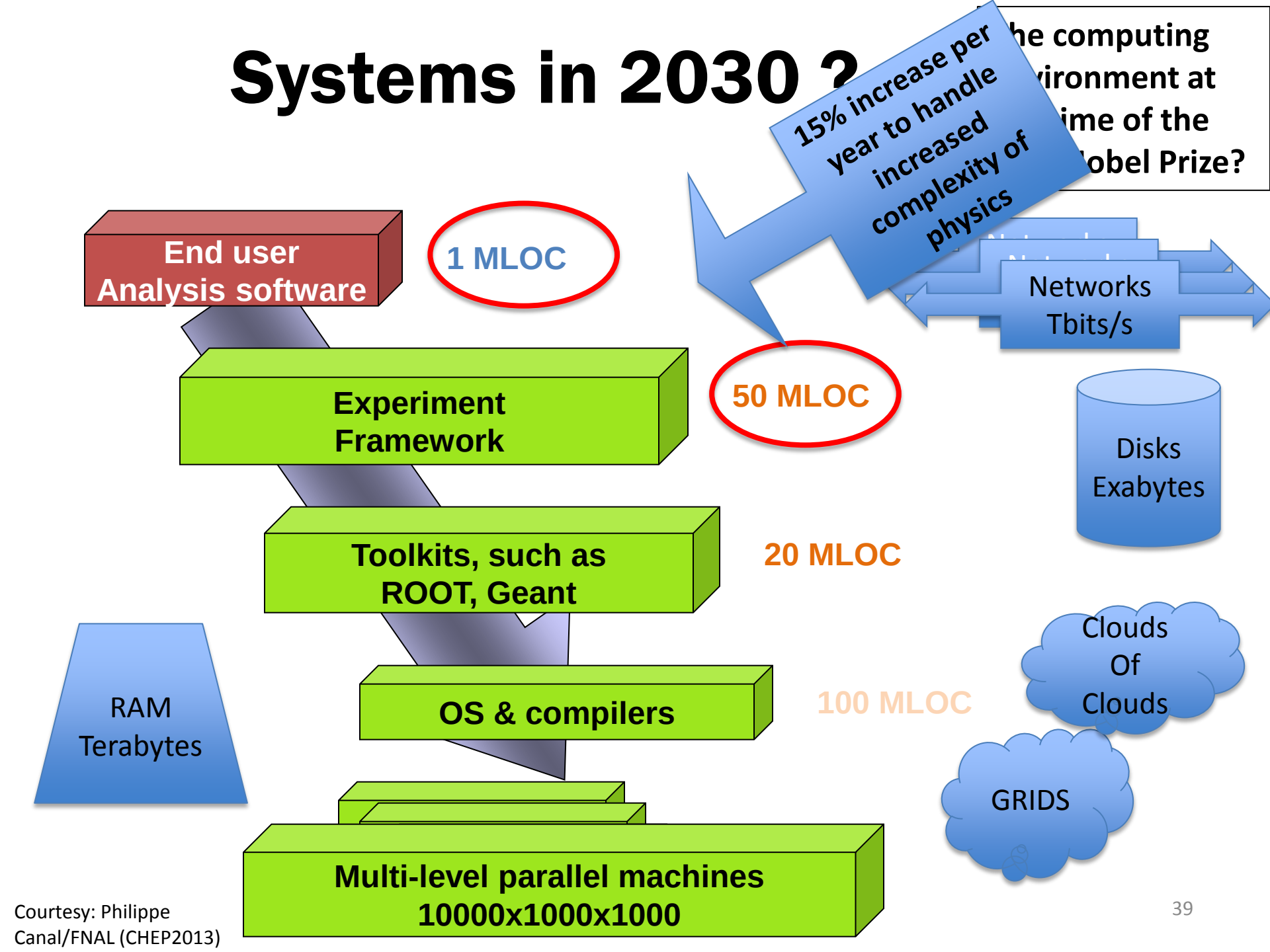
The computing environment at the time of Higgs's Nobel Prize (2013)



Courtesy: Philippe Canal/FNAL (CHEP2013)

Systems in 2030 ?

he computing
vironment at
ime of the
obel Prize?



Courtesy: Philippe Canal/FNAL (CHEP2013)

Conclusion

- Lots of exciting technology is on the 2020 horizon
- Things look good for "capacity exascale"
- Both offline and online physics computing will profit
  - Hopefully achieving better triggering, physics reconstruction, and analysis
- WLCG will reach millions of cores, exabytes of disk/tape storage, but will have to worry about power limitations (like everybody else)
- Large memories are seen as "opportunities"
- LHCb and ALICE are preparing exciting trigger systems based on high-speed networking, state-of-the art computing for 2019.
- Software complexity is daunting
- "Capacity-based" Big Data is where you want to be!